

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-185196

(43)Date of publication of application : 16.07.1996

(51)Int.Cl. G10L 3/00  
G10L 3/00

(21)Application number : 06-329161

(71)Applicant : SONY CORP

(22)Date of filing : 28.12.1994

(72)Inventor : MINAMINO KATSUKI  
ISHII KAZUO  
OGAWA HIROAKI

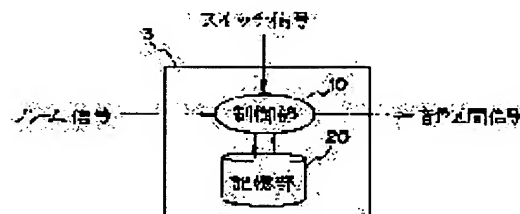
## (54) DEVICE FOR DETECTING SPEECH SECTION

## (57)Abstract:

PURPOSE: To detect a speech section without lacking a word head even if utterance is performed with timing at which a switch is not yet pressed by extracting only one speech section from a stored input speech signal.

CONSTITUTION: A speech section detecting device 3 has a control section 10 and a storage section 20 such as a memory. And the control section 10 always discriminates whether it is voice or voiceless for a speech signal made a digital signal by an A/D converter from a microphone, on the other hand, only one speech section is extracted from a speech signal stored in the memory part 20 in a wider range than a detection range specified by switch operation of a section specifying switch and outputted. Also, the memory part 20 always preserves an input speech signal of some fixed time.

Therefore, a speech section can be detected without lacking a word head even if utterance is performed with timing at which a switch is not yet pressed.



## LEGAL STATUS

[Date of request for examination]

11.01.2001

[Date of sending the examiner's decision of rejection]

30.06.2003

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-185196

(43) 公開日 平成8年(1996)7月16日

(51) Int.Cl. <sup>6</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 1 0 L 3/00	5 1 3 A			
	Z			
	5 7 1 G			

審査請求 未請求 請求項の数15 O L (全 13 頁)

(21) 出願番号 特願平6-329161

(22) 出願日 平成6年(1994)12月28日

(71) 出願人 000002185  
ソニー株式会社  
東京都品川区北品川6丁目7番35号

(72) 発明者 南野 活樹  
東京都品川区北品川6丁目7番35号 ソニー株式会社内

(72) 発明者 石井 和夫  
東京都品川区北品川6丁目7番35号 ソニー株式会社内

(72) 発明者 小川 浩明  
東京都品川区北品川6丁目7番35号 ソニー株式会社内

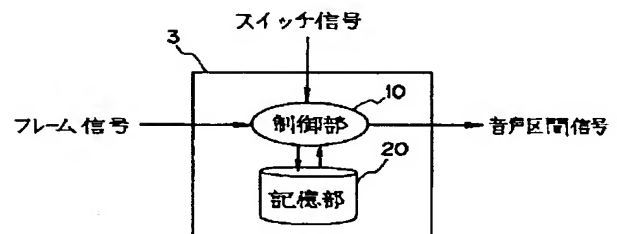
(74) 代理人 弁理士 小池 晃 (外2名)

(54) 【発明の名称】 音声区間検出装置

(57) 【要約】

【構成】 制御部10は、フレーム音声信号に対して、有音声と無音声とを常に判定しながら、一方でスイッチ操作によるスイッチ信号で指定される検出範囲よりも広い範囲で記憶部20に記憶されている上記音声信号から一つだけ音声区間を抽出し、出力する。

【効果】 スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出できる。



## 【特許請求の範囲】

【請求項1】 音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、  
上記入力音声信号を記憶する記憶手段と、  
上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有することを特徴とする音声区間検出装置。

【請求項2】 上記制御手段は、上記記憶手段から上記音声区間を抽出し、該音声区間を入力信号から一定時間遅らせて送信することを特徴とする請求項1記載の音声区間検出装置。

【請求項3】 上記制御手段は、上記入力された音声信号を周波数分析して信号パワースペクトルを求め、環境雑音のパワースペクトルの定数倍に対する該信号パワースペクトルの大小に応じて入力時の雑音を除去することを特徴とする請求項1記載の音声区間検出装置。

【請求項4】 上記制御手段は、上記信号パワースペクトルの積算値が所定のしきい値以上ならば有音声、未満ならば無音声と判定することを特徴とする請求項3記載の音声区間検出装置。

【請求項5】 上記制御手段は、上記有音声と判定される区間が所定長以上続いたときに、これを音声区間として検出し、その後に上記無音声と判定される区間が所定長以上続いたときに、上記音声区間が終了したと見なすことを特徴とする請求項4記載の音声区間検出装置。

【請求項6】 上記制御手段は、短い有音声部の後に無音声部が続く、その後再び有音声部が続くような場合、上記短い有音声部と上記無音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定することを特徴とする請求項5記載の音声区間検出装置。

【請求項7】 上記制御手段は、一定の長さの有音声部の後に、無音声部が続く、その後短い有音声が続くような場合、上記無音声部と上記短い有音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定することを特徴とする請求項5記載の音声区間検出装置。

【請求項8】 上記制御手段は、上記音声区間と判定された区間の前後にマージンを付加して、音声区間を引き延ばすことを特徴とする請求項5、6又は7記載の音声区間検出装置。

【請求項9】 上記制御手段は、上記有音声／無音声の判定のためのしきい値や、上記マージンのようなパラメータを環境雑音の平均エネルギーに応じて変動させることを特徴とする請求項4又は8記載の音声区間検出装置。

【請求項10】 上記制御手段は、上記パラメータと上記環境雑音の平均エネルギーとの関係を比例関係として予め決めておき、さらにそのパラメータの上限と下限も

2

決めておくことで、上記環境雑音の平均エネルギーから上記パラメータを決定することを特徴とする請求項9記載の音声区間検出装置。

【請求項11】 上記制御部は、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを無音声区間で更新することを特徴とする請求項9又は10記載の音声区間検出装置。

【請求項12】 上記制御部は、連続して無音声と判定され続け、かつ更新前の環境雑音の平均エネルギーから緩やかに変化するようなエネルギーを持つ区間においてのみ、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新することを特徴とする請求項11記載の音声区間検出装置。

【請求項13】 上記制御部は、一定時間以上有音声と判定され続けるような場合には、強制的に環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新することを特徴とする請求項9、10、11又は12記載の音声区間検出装置。

【請求項14】 上記制御部は、上記音声区間の抽出を常に行うことを特徴とする請求項1記載の音声区間検出装置。

【請求項15】 上記制御部は、上記環境雑音の平均エネルギー、上記環境雑音の平均パワースペクトル及び上記パラメータの更新を常に行うことを特徴とする請求項9、10、11、12又は13記載の音声区間検出装置。

## 【発明の詳細な説明】

## 【0001】

【産業上の利用分野】 本発明は、入力された音声信号から、有音声部と無音声部を識別し、スイッチ操作と連動させて必要な音声区間のみを検出する音声区間検出装置に関する。

## 【0002】

【従来の技術】 従来、音声区間を検出するには、音声信号のパワーがあるしきい値を越えれば有音声部、そのしきい値を越えなければ無音声部と判定し、ある一定長以上の有音声部が得られたときに、これを音声区間として検出するという方法がある。また、有音声部と無音声部の判別に、零交差数を用いたり、周波数分解したスペクトルパワーを用いたりする方法、さらに、これらを組み合わせさせて判別する方法もある。また、環境雑音の平均エネルギーに応じてしきい値を変動させたり、環境雑音の平均スペクトルと入力信号のスペクトルとの類似性や相関などを調べることで、雑音環境下でも音声区間が頑強に検出できるようにした方法等もある。これらの方法をそのまま音声認識システムや音声通信システムの前処理として適用した場合、うまく音声区間を検出することができれば、後段の処理が正常に動作することが期待できるわけだが、実際には、その音声区間の前後に発生される不要な音声も検出されてしまい、システム全体の動作

3

に悪影響を及ぼすという問題が生じる。例えば、地名を認識する音声認識システムでは、話者が「しながわ」と発声すれば、その発声に対して認識を行うようにしたいわけだが、その発声の前に「えーっと、」などと発声した場合にはこれも音声区間として検出されてしまい、その発声に対して認識が行われることになってしまう。同様に、「しながわ」と発声した後しばらくしてから「・・・だよね？」などと横にいる話し手に話しかけた場合、話者はシステムに話したつもりはなくても、やはりこれが音声区間として検出されてしまい、その発声に対しても認識が行われることになってしまう。

【0003】これらの問題は、音声認識システムがユーザの発声した音声の中で、必要な音声区間だけを選ぶことができないことに起因する。

【0004】そこで、スイッチを付け、必要な音声区間をユーザが直接指定するようにすることで、この問題を解決する方法が考えられる。一つは、スイッチが押されている間を音声区間として抽出する方法、一つは、スイッチが一度押された後、一定時間の間に発声された音声区間を前述の有音声と無音声の判別方法を用いて抽出する方法などである。

【0005】

【発明が解決しようとする課題】ところで、これらの方法においては、ユーザが発声より少しでも遅れてスイッチを押した場合に、抽出すべき音声区間の最初の部分が欠けてしまうという問題が生じてしまう。また、環境雑音の平均エネルギーに応じてしきい値を変動させたり、環境雑音の平均スペクトルを用いるなどするような雑音環境下にも適応できる区間検出方法を用いた場合、従来は、その環境雑音を抽出するためにスイッチを押してから一定時間環境雑音を観測するというを行うため、その間ユーザは話せないなどの煩わしさを伴うという問題があった。

【0006】本発明は、上記実情に鑑みてなされたものであり、ユーザにスイッチを押すタイミングと発声のタイミングを意識させるような煩わしさを不要とさせ、必要な音声区間だけをうまく検出し、かつ雑音環境下でも安定した音声区間検出を実現できる音声区間検出装置の提供を目的とする。

【0007】

【課題を解決するための手段】本発明に係る音声区間検出装置は、音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、上記入力音声信号を一定時間分記憶する記憶手段と、上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有することにより上記課題を解決する。

【0008】この場合、上記制御手段は、上記記憶手段から上記音声区間を抽出し、該音声区間を入力信号から

4

一定時間遅らせて送信する。

【0009】また、上記制御手段は、上記入力された音声信号を周波数分析して信号パワースペクトルを求め、環境雑音のパワースペクトルの定数倍に対する該信号パワースペクトルの大小に応じて入力時の雑音を除去する。

【0010】また、上記制御手段は、上記信号パワースペクトルの積算値が所定のしきい値以上ならば有音声、未満ならば無音声と判定する。

10 【0011】また、上記制御手段は、上記有音声と判定される区間が所定長以上続いたときに、これを音声区間として検出し、その後上記無音声と判定される区間が所定長以上続いたときに、上記音声区間が終了したと見なす。

【0012】また、上記制御手段は、短い有音声部の後に無音声部が続く、その後再び有音声部が続くような場合、上記短い有音声部と上記無音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定する。

20 【0013】また、上記制御手段は、一定の長さの有音声部の後に、無音声部が続く、その後短い有音声が続くような場合、上記無音声部と上記短い有音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定する。

【0014】また、上記制御手段は、上記音声区間と判定された区間の前後にマージンを付加して、音声区間を引き延ばす。

30 【0015】また、上記制御手段は、上記有音声／無音声の判定のためのしきい値や、上記マージンのようなパラメータを環境雑音の平均エネルギーに応じて変動させる。

【0016】また、上記制御手段は、上記パラメータと上記環境雑音の平均エネルギーとの関係を比例関係として予め決めておき、さらにそのパラメータの上限と下限も決めておくことで、上記環境雑音の平均エネルギーから上記パラメータを決定する。

【0017】また、上記制御部は、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを無音声区間で更新する。

40 【0018】また、上記制御部は、連続して無音声と判定され続け、かつ更新前の環境雑音の平均エネルギーから緩やかに変化するようなエネルギーを持つ区間においてのみ、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新する。

【0019】また、上記制御部は、一定時間以上有音声と判定され続けるような場合には、強制的に環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新する。

50 【0020】また、上記制御部は、上記音声区間の抽出を常に行う。

5

【0021】また、上記制御部は、上記環境雑音の平均エネルギー、上記環境雑音の平均パワースペクトル及び上記パラメータの更新を常に行う。

【0022】

【作用】有音声と無音声の判別による音声区間判定を常に行いながら、その判定結果と一定時間前までの入力信号を記憶しておくことで、スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出することが可能と成る。又、一つのスイッチ指定区間に対して一つの音声区間だけを検出し送信するようにしたことで、話者が発声の少し前や、少し後に関係のない言葉を発声したとしても、これを音声区間として検出することがなくなるため、例えば音声認識装置に適用した場合などに誤動作を起こしにくくなる。

【0023】さらに、入力信号を周波数分析して得られるパワースペクトルを用いること、及び、環境雑音の平均エネルギーに適応させてパラメータを更新することなどを行うことで、環境雑音下においても安定した音声区間検出が行えるようにすると共に、その適応化を音声区間の判定と平行して常に行っておくようにすることで、スイッチを押してから雑音環境の観測を行う必要が特になくなるため、ユーザが発声するのを待たされることもなくなる。つまり、ユーザは、単純にスイッチを押せば一回だけ発話することができるということだけを覚えておけばよく、その発話とスイッチのタイミングをそれほど意識しなくてよいものとなる。しかも、システムは、必要な音声区間だけを検出することが可能となる。

【0024】そして、この音声区間検出装置を音声認識システムや音声通信システムに組み込めば、入力信号のなかから認識処理を行うべき音声区間や送信すべき音声区間を抽出することができるようになる。

【0025】

【実施例】以下、本発明に係る音声区間検出装置の実施例について説明する。この実施例となる音声区間検出装置は、話者にスイッチを押すタイミングと発声のタイミングとをあまり意識させることなく、必要な音声区間を検出できる。この音声区間検出装置は、例えば図1に示すような音声認識システムに適用され、検出した音声区間は音声認識のために使われる。まず、本実施例の音声区間検出装置を説明する前に、図1に示す音声認識システムについて説明する。

【0026】マイクロホン1で收音された音声信号は、A/D変換器2でデジタル信号に変換され、本実施例の音声区間検出装置3に供給される。この音声区間検出装置3には、話者が押した区間指定スイッチ4からのオン、オフ信号も供給される。そして、音声区間検出装置3は、上記デジタル信号と、上記オン、オフ信号を用いることで、必要な音声区間の音声信号を検出し、これを音声認識部5に送る。音声認識部5は、音声区間検出

6

装置3が検出した音声信号に対して認識処理、すなわち音響分析やベクトル量子化などによる特徴量の抽出と、ダイナミックプログラミング（以下、DPという。）マッチングやHMMなどによるスコア計算を行ない、その認識結果を出力する。このような音声認識システムは、さまざまな機器のコントロールを音声で行う場合や、キーボードなどに変わる入力手段の一つとして、広く用いられることが期待できる。

【0027】このような音声認識システムにおいて、話者にあまり煩わしさを与えることなく、必要な音声区間を抽出することが重要となる。このため、本実施例となる音声区間検出装置3が必要となる。

【0028】この音声区間検出装置3は、図2に示すように、制御部10と、例えばメモリのような記憶部20とを有して成る。制御部10は、上記図1に示したA/D変換器2でデジタル信号とされたマイクロホン1からの音声信号に対して、有音声と無音声とを常に判定しながら、一方で区間指定スイッチ4のスイッチ操作で指定される検出範囲よりも広い範囲で記憶部20に記憶されている上記音声信号から一つだけ音声区間を抽出し、出力する。また、制御部10は、環境雑音の観測、内部パラメータの更新なども行うようにする。記憶部20は、ある一定時間分の入力音声信号を常に保存しておく。

【0029】制御部10に入力されるA/D変換器2からのデジタル入力音声信号は、サンプリングされたデータであり、例えば128サンプルをまとめて1フレームとされ、1フレームずつ順に送られて来るものとする。なお、以下では、時刻 $t$ に制御部10に入力されるフレーム信号を $F_t$ と表す。

【0030】この制御部10は、図3に示すフローチャートに基づいた動作を行う。

【0031】まず、このフローチャートが開始されると、この制御部10は、ステップS1に示すように、上記フレーム信号 $F_t$ を記憶部20に保存する。ここで、この記憶部20には、常に $k$ フレーム前までの入力信号を記憶しておくようにする。すなわち、記憶部20は、時刻 $t$ には $k$ 時刻前までのフレーム信号 $F_t$ 、 $F_{t-1}$ 、 $F_{t-2}$ ・・・ $F_{t-k}$ を記憶している。

【0032】次に、この制御部10は、ステップS2からステップS5までに示すような動作を行い、時刻 $t$ の入力フレーム信号 $F_t$ に音声が含まれるかどうか、つまり有音声か無音声かを判別する。ステップS2からステップS5までに示す動作は、周波数分析を用いる方法を適用している。

【0033】ステップS2では、入力されたフレーム信号 $F_t$ の周波数分析を行う。具体的には、入力されたサンプル信号に、ハミングウィンドウをかけ、高速フーリエ変換（以下、FFTという。）を施すことで、パワースペクトルを求めたり、バンドパスフィルタを用いて各

7

帯域毎のパワースペクトルを抽出する。

【0034】ステップS3では、環境雑音の平均パワースペクトルを用いて、雑音除去を行う。具体的に、雑音除去後のパワースペクトル $X(\omega)$ は、入力信号のパワースペクトルを $S(\omega)$ 、環境雑音の推定パワースペクトルを $N(\omega)$ 、オフセットを $R$ 、環境雑音の推定パワースペクトル $N(\omega)$ にかける重み係数を $\alpha$ とすると、

【0035】

【数1】

$$X(\omega) = \begin{cases} S(\omega) + R & S(\omega) \geq \alpha \cdot N(\omega) \\ R & S(\omega) < \alpha \cdot N(\omega) \end{cases} \quad \dots(1)$$

【0036】のように求められる。したがって、この制御部10は、ある帯域 $\omega$ において、入力信号のパワースペクトル $S(\omega)$ が環境雑音の推定パワースペクトル $N(\omega)$ の $\alpha$ 倍以上ならば、その帯域 $\omega$ には雑音以外のパワースペクトルが多く含まれていると見なしオフセット $R$ を付加して上記入力信号のパワースペクトル $S(\omega)$ をそのまま残し、逆に、もし上記入力信号のパワースペクトル $S(\omega)$ が上記環境雑音の推定パワースペクトル $N(\omega)$ の $\alpha$ 倍よりも小さいならば、その帯域には雑音しか含まれていないと見なしオフセット $R$ のみとすることで、雑音のパワースペクトルを除去している。これにより、環境雑音と周波数成分の異なる信号がある程度の大きさで入力されれば、その周波数成分の存在する帯域において、入力信号のパワースペクトル $S(\omega)$ が残ることになる。例えば、静かな環境、ファンノイズの環境、自動車走行雑音の環境など、さまざまな雑音環境下において音声が発声された場合、環境雑音が多少大ききときでも、音声帯域と環境雑音の帯域の相違から、雑音除去後のパワースペクトル $X(\omega)$ の音声帯域における入力信号のパワースペクトル $S(\omega)$ は除去されずに残ることになる。

【0037】そこで、制御部10は、ステップS4に示すように、ステップS3で得られたパワースペクトル $X(\omega)$ からエネルギー $E$ を、

【0038】

【数2】

$$E = 10 \log_{10} \left( \sum_{\omega} X(\omega) \right) \quad \dots(2)$$

【0039】のように求めた後、このエネルギー $E$ があるしきい値 $r$ より大きいかな否かを判定し、ステップS5に示すように有音声／無音声の判定を行っている。上記(2)式のエネルギー $E$ が上記しきい値 $r$ 以上ならば有音声、上記しきい値未満ならば無音声と判定する。ここで、しきい値 $r$ としては、例えば、

【0040】

【数3】

$$r = 10 \log_{10} \left( \sum_{\omega} R \right) + \delta \quad \dots(3)$$

【0041】のような一定値を用いたり、あるいは後述するような、環境雑音の平均エネルギーに適應して変動させた値を用いればよい。ただし、 $\delta$ は定数である。

【0042】なお、ここでは、有音声／無音声判定部の判定方法として、周波数分析により得たパワースペクトルを用いる方法を述べたが、これに代わる方法として、従来のパワーや零交差数などによる判定方法を用いてもよい。

【0043】次に、制御部10は、ステップS6に示すように、音声区間の判定を行う。音声区間の判定は、上述したような有音声と無音声の判定に基づき、有音声部が最低 $m$ フレーム（例えば、30フレーム）以上続く場合の有音声部の始端から終端までを、基本的に音声区間として判定するようにする。また、 $n$ フレーム（例えば25フレーム）以上の無音声部が続けば、これを無音声区間と判定し、上記音声区間が終了したと見なす。

【0044】この時、音声区間の始端をできるだけ早く検出するため、図4の(A)に示すように、フレーム $F_t$ が入力されたときに、フレーム $F_{t-m}$ が音声区間のフレームであるかどうかを判定し、その結果を記憶部20に書き込むようにする。すなわち、時刻 $t$ に右端のフレーム $F_t$ が入力されたときに、 $m$ 時刻前のフレーム $F_{t-m}$ が音声区間かどうかを判定し、その結果を記憶部20に書き込ませる。図4の(A)の斜線部分は、有音声部を表す。また、後述する図4の(B)乃至図4の(G)において、示される斜線部分も有音声部を表し、それを除く部分は無音声部を表す。

【0045】以下、図4の(B)乃至図4の(G)を参照しながらステップS6の音声区間判定の動作の詳細を説明する。図4の(B)は、音声が入力されておらず、無音声区間中にさらに無音声フレームが入力された場合であり、 $F_{t-m}$ は無音声区間として判定される。図4の(C)は、音声が入力され始め、有音声フレームが数フレーム連続して入力された場合であるが、まだ $m$ フレーム以上の有音声部が続いていないので、やはり $F_{t-m}$ は無音声区間として判定される。図4の(D)は、フレーム $F_t$ が有音声部として判定され、有音声部がちょうど $m$ フレーム続いたときを示し、このときフレーム $F_{t-m}$ が音声区間の始まりとして検出される。その後、図4の(E)のように、しばらく入力フレームが有音声と判定された場合、フレーム $F_{t-m}$ は音声区間として判定され続ける。しばらくして、音声の入力が終了すれば、図4の(F)のように無音声フレームが入力され始める。しかし、フレーム $F_{t-m}$ はまだ音声区間として判定されたままとなる。そして、図4の(G)のように $n$ フレームの無音声フレームが続けて入力された場合に、音声区間の終了が検出される。ただし、図4の(G)は $n \leq m$ の

場合を示しており、この場合は、その後 $m-n$ フレームが音声区間として判定されることになる。そしてしばらくして、再び図4の(B)に示す状態に戻る。

【0046】また、 $n > m$ の場合は、図4の(C)が少し変わり、音声の入力が終了し、連続して $m$ フレーム無音声フレームが入力されても、フレーム $F_{t-m}$ は音声区間として判定されたままとなる。そこで、 $n$ フレームの連続した無音声フレームが入力された時点で、 $F_{t-m}$ を無音声区間と判定し、同時に、その $n-m$ フレーム前までの音声区間と判定されたフレーム $F_{t-m-1}$ 、 $F_{t-m-2}$ 、 $\dots$ 、 $F_{t-n}$ の判定結果を無音声区間と書き換えるようにする。

【0047】なお、音声区間の判定結果として、音声区間の始端や終端が検出された時に、どのフレームが始端であるとか、どのフレームが終端であるとかの情報も記憶部20に書き込むようにするものとする。

【0048】このステップS6に示すような音声区間の判定を制御部10が行う場合、連続して有音声部と判定されるような図5の(A)に示す音声信号に対しては問題ないが、図5の(B)に示すような音声区間Vの前部において無音声部Uに分離された短い有音声部vが存在する場合や、図5の(C)に示すように音声区間V終了後に無音声部Uが一旦入力され、続いて短い有音声部vが入力された場合などに、これらを音声区間として含めるかどうかという問題に対処できない。このような前後に付加された短い有音声部vは、音声区間として含めた方が良いものもあれば、誤って有音声部と判定された非定常ノイズなど、音声区間には含めない方が良いものもある。

【0049】そこで、ステップS6の音声区間判定時に、以下に述べるような処理を追加する。まず、問題となるのは、 $m$ フレーム未満しか連続しない有音声部の後に無音声部が入力され、その後再び有音声部が入力された場合の処理である。そこで、この $m$ フレーム未満の有音声部のフレーム数とその後に続く無音声部のフレーム数をカウントするようにし、 $m$ フレーム未満の有音声部の後に、無音声フレームが連続して入力された場合に、音声区間と見なすか見なさないかを図6の(A)のような関係に基づいて判定する。すなわち、もし、 $m_{pre}$ フレーム(例えば3フレーム)未満の有音声部vの後に $n_{pre}$ フレーム以上の無音声部Uが入力された場合は、その $m_{pre}$ フレーム未満の有音声部vは音声区間から除去するものとする。ただし、 $m_{pre} < m$ とする。また、 $x$ フレーム( $m_{pre}$ フレーム以上 $m$ フレーム未満)の有音声部の後に、

【0050】

【数4】

$$n_{pre} + \frac{n - n_{pre}}{m - m_{pre}} x \quad \dots (4)$$

【0051】フレーム以上の無音声部が連続して入力されれば、その $x$ フレームの有音声部も音声区間から除去する。つまり、図6の(A)に示すように、横軸を有音声部の入力フレーム数、縦軸をその後に続く無音声部のフレーム数としたとき、斜線部の関係の場合は、その有音声部を音声区間を含めず、斜線部より下の関係の場合は、その有音声部との間にはさまった無音声部を音声区間を含めるようにする。

【0052】以上のような処理を追加することにより、図5(B)に示されるような音声区間Vの前部の短い有音声部vの扱いとして、挿入された無音声部Uが短い時には該短い有音声部vを音声区間を含め、長い時には該短い有音声部vを音声区間には含めないようにすることが可能となる。

【0053】同様に、図5の(C)のような音声区間Vの最後に無音声部Uで分離された短い有音声部vが存在する場合の処理も同じように行う。すなわち、図6の

(B)に示されるように、 $m$ フレーム以上の有音声部の後に発生した無音声部のフレーム数(縦軸)とその後に表れる $m$ フレーム未満の有音声部のフレーム数(横軸)をカウントするようにし、無音声フレームが連続して入力された後に、 $m$ フレーム未満の有音声部vが入力された場合に、音声区間と見なすか見なさないかを判定している。つまり、図6の(B)に示す斜線部のような関係であれば、その最後の有音声部分は音声区間から除去し、斜線部の下のような関係であればその最後の有音声部分との間に挟まった無音声部を音声区間を含めるようにする。すなわち、 $n_{post}$ フレーム以上の無音声部Uが入力された後に、 $m_{post}$ フレーム(例えば3フレーム)未満の有音声部vが入力された場合は、その $m_{post}$ フレーム未満の有音声部vは音声区間から除去する。ただし、 $m_{post} < m$ とする。また、 $x$ フレーム( $m_{post}$ フレーム以上 $m$ フレーム未満)の有音声部の前に、上記(4)式と同様の式、すなわち上記(4)式の $n_{pre}$ を $n_{post}$ 、 $m_{pre}$ を $m_{post}$ と変更した式によって表されるフレーム以上の無音声部が連続して入力されていれば、その $x$ フレームの有音声部も音声区間から除去する。

【0054】次に、制御部10は、ステップS7に示すように、ステップS6で音声区間と判定されたフレームの前後に、さらに音声区間としてのマージンを付加し、実際の音声区間より少し長めの音声区間を抽出する。これは、ステップS6の検出誤り、すなわち音声区間の始端が遅れて検出されてしまったり、終端が速く検出されてしまうのを防ぐために行う。

【0055】例えば、図4の(D)に示すように、音声区間の始端が検出された時点、すなわち、時刻 $t$ にフレーム $F_t$ が入力され、その $m$ 時刻前のフレーム $F_{t-m}$ が音声区間の始まりとして検出されたときに、記憶部20にフレーム $F_{t-m}$ が音声区間であると書き込むと同時に、

そこからさらに $p$ フレーム前までのフレームに対しても

11

音声区間であったことを追加して書き込むようにすればよい。

【0056】同様に、図4の(E)のような音声区間の終了が検出され、その終了フレームの判定結果が記憶部20に書き込まれてから、さらにqフレーム後まで音声区間とみなして、記憶部20に書き込み続けるようにする。以上のようにして、ステップS6において、判定された音声区間のフレームの前にpフレーム、後ろにqフレームのマージンを付加したものが最終的な音声区間として記憶部20に記憶されていくことになる。その際、音声区間の始端や終端の情報も、対応したものに変更しておくようにする。ここで、マージンp、qは一定値を用いたり、あるいは後述するような、環境雑音の平均エネルギーに応じて変動させた値を用いてもよい。

【0057】次に、制御部10は、ステップS8に示すように、音声区間の判定のためのパラメータを環境の変化に応じて更新する。音声区間の判定のためのパラメータ、すなわち環境雑音の推定パワースペクトル $N(\omega)$ や環境雑音の平均エネルギー $E_n$ 、さらに、有音声・無音声の判定のためのしきい値 $r$ や音声区間のマージンp、qは、環境の変化に応じて変動させることが、耐雑音性能の向上のために必要となってくる。そこで、これらのパラメータの更新は毎フレーム行うことにし、一つのフレームの処理が終了した時点で、次フレームの処理のために新しく更新を行うようにする。

【0058】まず、環境雑音の推定パワースペクトル $N(\omega)$ と環境雑音の平均エネルギー $E_n$ の更新について説明する。これは基本的には、音声区間外において、環境雑音の推定パワースペクトル $N(\omega)$ や環境雑音の平均エネルギー $E_n$ を平均化したものを求めるようにする。その求めかたとしては、まず、入力フレームから求められるパワースペクトル $S(\omega)$ からエネルギー $E_s$ を、

【0059】

【数5】

$$E_s = \sum_{\omega} S(\omega) \quad \dots(5)$$

【0060】のように求め、そして、前フレームにおいて求められた上記環境雑音の推定パワースペクトル $N(\omega)$ と上記環境雑音の平均エネルギー $E_n$ を用いて、例えば、

【0061】

【数6】

$$N(\omega) := \frac{(h-1)N(\omega) + S(\omega)}{h} \quad \dots(6)$$

【0062】

【数7】

$$E_n := \frac{(h-1)E_n + E_s}{h} \quad \dots(7)$$

【0063】のように更新する。

12

【0064】ここで、h(例えば、20)は重み係数であり、更新前の値に $(h-1)/h$ 、入力フレームの値に $1/h$ の重みをかけて加え合わせることによって、時間的に新しいフレームに重みをおいて平均化を行った結果が得られることになる。

【0065】なお、上記の更新は、音声区間外のフレームにおいてのみ行うものとし、例えば、図4の(B)や、図4の(G)のように、nフレーム前から連続して無音声と判定されつづけ、しかも前フレームで求められた環境雑音の平均エネルギー $E_n$ に比べて入力フレームのエネルギー $E_s$ が急激に大きくなり過ぎないように入力フレーム、例えば、 $E_n$ と $E_s$ を比較したとき、 $E_n/E_s > 0.5 \quad \dots(8)$

を満たすような入力フレームにおいてのみ更新を行うようにする。つまり、有音声と判定されるフレームや、既に求めてある環境雑音の平均エネルギーから大きく離れたエネルギーをもつようなフレームにおいては、環境雑音のパラメータ更新を行わないようにすることで、ある程度安定した環境雑音の推定パワースペクトル $N(\omega)$ や環境雑音の平均エネルギー $E_n$ が得られることになる。

【0066】このようにして得られた環境雑音の平均エネルギー $E_n$ に適応させて、他のパラメータを更新する方法に関して説明する。あるパラメータPを、環境雑音の平均エネルギー $E_n$ が大きなどときには大きく、 $E_n$ が小さいときには小さくしたい場合、簡単には、パラメータPを環境雑音の平均エネルギー $E_n$ やその対数に比例させて変動させる方法が考えられる。この際、パラメータPの変動の範囲を制限するため、パラメータPの上限下限を設定する。これを示したのが図7の(A)である。横軸が環境雑音の平均対数エネルギーHであり、この平均対数エネルギーHは、

【0067】

【数8】

$$H = 10 \log_{10} E_n \quad [dB] \quad \dots(9)$$

【0068】のように求められる。また、縦軸がパラメータPである。 $P_{max}$ 、 $P_{min}$ は、パラメータPの下限上限を示している。また、比例定数は、例えば、パラメータPを環境雑音が $H_a$  [dB]の場合に $P_a$ 、 $H_b$  [dB]の場合に $P_b$ に設定したい場合、

【0069】

【数9】

$$\frac{P_b - P_a}{H_b - H_a} \quad \dots(10)$$

【0070】のようにして求めることができる。このようにして、図7の(A)のような関係を予め決めておけば、環境雑音の平均エネルギー $E_n$ からパラメータPを求めることが可能となる。同様に、パラメータPを、環境雑音の平均エネルギー $E_n$ が大きいときに小さく、環



13

環境雑音の平均エネルギー  $E_n$  が小さいときに大きくしたい場合には、図 7 の (B) のような関係に基づいて求めれば良い。比例定数はやはり上記 (10) 式で求められる。

【0071】このような環境雑音の平均エネルギー  $E_n$  に適応させたパラメータ  $P$  の更新方法は、例えば、音声区間の前後に付加するマージン  $p$ 、 $q$  のように、環境雑音が大きき場合には音声区間の判定の精度が悪くなるので大きくし、環境雑音が小さい場合には音声区間の判定精度が良いので小さくしたいというときなどに用いることができる。また、上記 (1) 式の重み係数  $\alpha$  や、有音声・無音声の判定のためのしきい値  $r$  などを、環境雑音が大きき時には大きく、小さいときには小さくしたい場合、あるいは逆に、環境雑音が大ききときには小さく、小さいときには大きくしたい場合などに適用することができる。

【0072】以上、環境雑音の平均エネルギー  $E_n$  の更新方法と、該エネルギー  $E_n$  に応じた他のパラメータ  $P$  の更新方法に関して述べた。基本的には、あるパラメータ  $P$  を環境に適応させて変動させたい場合、パラメータ  $P$  と環境雑音の平均エネルギー  $E_n$  の対応関係である

$P = f(E_n)$  を予め決めておき、環境雑音の平均エネルギー  $E_n$  からパラメータ  $P$  を求めるようにすればよい。

【0073】なお、パラメータ  $P$  を一定にすることは、 $P_{max} = P_{min} = \text{constant}$  とすることに対応する。

【0074】次に、制御部 10 は、ステップ S 8 で示したパラメータ更新の例外として、音声区間が一定以上長く続くような場合に、ステップ S 9 に示すように強制的なパラメータ更新を行う。

【0075】上記ステップ S 8 のパラメータ更新において、環境雑音の推定パワースペクトル  $N(\omega)$  や環境雑音の平均エネルギー  $E_n$  は、音声区間外のフレームにおいてのみ更新を行うものとしたが、例外として、音声区間がある一定以上長く続くような場合には、強制的に更新を行うようにしておく。つまり、図 4 の (E) のような状態が長く続き、音声区間として判定されるフレームが  $Q$  (例えば 500) フレーム以上続いた場合に、上記  $N(\omega)$  や上記  $E_n$  を強制的に更新するようにする。これは、環境雑音が急激に大きくなった場合などに、誤って有音声と判定し続け、しかも、環境雑音の推定パワースペクトル  $N(\omega)$  や環境雑音の平均エネルギー  $E_n$  がこれに追従できないという問題が生じるのを避けるためである。このような処理を付加しておくことで、環境雑音の変動に対してさらに強くすることが可能となる。

【0076】次に、制御部 10 は、ステップ S 10 に示すように、最終的に音声区間を検出し、送信を行う。

【0077】上述したステップ S 9 までの処理が終了したとき、図 2 の記憶部 20 には入力フレーム  $F_t$  から  $k$  フレーム前までのフレーム信号と、その時点までの判定

14

結果、すなわち、 $F_{t-m}$ 、 $F_{t-m-1}$ 、 $\dots$ 、 $F_{t-k}$  に対して付けられた音声区間の判定結果、及び音声区間の始端・終端情報が記憶されていることになる。ここで、 $k$  は音声区間として判定するために最低必要なフレーム数  $m$  に、音声区間の始端を前にずらすためのマージン  $p$  の最大値を加えたものより大きく、かつ無音声の区間を判定するために最低必要なフレーム数  $n$  よりも大きいものとする。これは、音声区間の判定が行われるのが入力されてから  $m$  フレーム後であり、しかもマージンが付加される場合は、さらに  $p$  フレーム後となるため、そのフレームの判定が確定するのは入力されてから  $m$  に  $p$  の最大値を加えたフレーム分だけ遅れてからとなること、及び、音声区間の終了が検知できるのは入力されてから  $n$  フレーム遅れてからとなることから要求される。

【0078】そして、制御部 10 は、この記憶部 20 から、ステップ S 10 に示すように、必要な音声区間を検出して出力する。図 1 の区間指定スイッチ 4 が押された場合、図 2 の制御部 10 に対してスイッチ信号が送られて来る。これには、ほとんど遅延がない。この区間指定スイッチ 4 のスイッチ信号の送られてくる様子を示したのが、図 8 の (A) である。横軸が時間、縦軸がスイッチ信号であり、区間指定スイッチ 4 が押されている間は、“1”、押されていないときは“0”の信号が送られて来るものとする。これに対して、入力フレーム信号は図 8 の (B) に示すように  $\Delta$  時間毎に送られてくる。つまり、区間指定スイッチ 4 が押され出した時点  $t_s$  と区間指定スイッチ 4 が離された時点  $t_e$  は、図 8 の (C) に示されるように、あるフレームが入力されてから次のフレームが入力されるまでの  $\Delta$  の間に検知されるはずである。

【0079】そこで、区間指定スイッチ 4 の押され出した時点  $t_s$  の直後に入力されたフレーム (図 8 の (C) の場合  $F_{t+1}$ ) を区間指定スイッチ 4 のオンの起点とし、区間指定スイッチ 4 が離された時点  $t_e$  の直後のフレーム (図 8 の (C) の場合  $F_{t+4}$ ) を区間指定スイッチ 4 のオフの起点とする。そして、図 9 の (A) に示すように、区間指定スイッチ 4 のオンの起点の  $k$  フレーム前から、区間指定スイッチ 4 のオフの起点の  $l$  フレーム後ろまでの区間をスイッチ指定区間とする。このとき、各フレーム信号が入力される度に、区間指定スイッチ 4 のオン、オフの起点となるかどうかを調べることができ、これを基にして記憶されている  $k$  フレーム前から入力フレームまでの信号がスイッチ指定区間に含まれるか否かを判定できるので、その判定結果も記憶部 20 に書き込むことにする。この際、スイッチ指定区間の始端と終端の情報も記憶するようにしておく。そして、このスイッチ指定区間の中に、音声区間の始端が検出されれば、その音声区間の始端から終端までを必要な音声区間として図 10 の (A) に示す範囲で検出し、これを送信する。ここで、音声区間がスイッチ指定区間終了後も続

15

くような場合でも、図10の(B)に示すように、その音声区間の終了までは送信するものとする。

【0080】送信の仕方は、図11に示すようにする。すなわち、送信すべき音声区間の信号を入力させてから、k時刻遅れて送信する。時刻tにおいては、図2の記憶部20に記憶されているk時刻前のフレーム $F_{t-k}$ を常に着目し、そのフレームが送信すべき音声区間のフレームであるときのみ送信し、そうでないときは送信しない。こうすることによって、送信すべき音声区間のフレームは、入力フレーム $F_t$ に同期して、k時刻遅れて送信されることになるわけである。

【0081】次に、送信すべきか否かの判断について述べる。時刻tにおいて着目するのは、k時刻前のフレーム $F_{t-k}$ である。この時点において、記憶部20には、フレーム信号 $F_{t-k}$ が記憶されていると同時に、これがスイッチ指定区間であるかどうか、音声区間であるかどうかにも既に判定され記憶されていることになる。また、音声区間であれば、音声区間の始端・終端であるかどうかにも記憶されていることになる。そこで、図12に示すような状態遷移図にしたがって、着目フレーム $F_{t-k}$ がどの状態からどの状態に遷移するかを調べ、その状態遷移に応じて送信すべきかどうかを決定するようにする。

【0082】先ず最初、状態(i)から始める。そして、区間指定スイッチ4のオンの起点が検知されるまで、状態(i)で自己遷移する。もし、区間指定スイッチ4のオンの起点が検知されれば、着目フレーム $F_{t-k}$ が音声区間の始端のフレームかどうかを調べ、NOならば状態(ii)に、YESならば状態(iii)に遷移する。状態(ii)に遷移した後は、着目フレーム $F_{t-k}$ が音声区間の始端のフレームかどうかを調べ、YESならば状態(iii)へ遷移する。着目フレームが音声区間の始端でない場合は、さらに着目フレームがスイッチ指定区間の終了フレームかどうかを調べ、NOならば状態(ii)で自己遷移し、YESならば再び状態(i)に戻る。状態(iii)においては、着目フレームが音声区間の終端フレームかどうかを調べ、NOならば状態(iii)で自己遷移する。もし、着目フレームが音声区間の終端フレームとなった場合は、その時点からk時刻前まで、つまりフレーム $F_{t-k-1}$ から $F_t$ が入力された間に区間指定スイッチ4のオンの起点が検知されたかどうかを調べ、YESならば状態(ii)に、NOならば状態(i)に遷移する。そして、このような状態遷移において、状態(iii)への遷移時と状態(iii)からの遷移時、すなわち、状態(i)から状態(iii)、状態(ii)から状態(iii)、状態(iii)における自己遷移、状態(iii)から状態(i)、状態(iii)から状態(ii)の遷移時において、着目フレーム $F_{t-k}$ を送信するようにする。逆に、それ以外の遷移時には、送信を行わないようにする。

【0083】以上のような処理により、着目フレームがスイッチ指定区間に入った後、音声区間の始端が検知さ

16

れてから終端が検知されるまで送信が行われるようになる。もし、スイッチ指定区間が終了するまでに、音声区間の始端が検知されないような場合は、何も送信されないままとなる。(状態(ii)から状態(i))。また、区間指定スイッチ4のオンの起点が検出されない限り、状態(i)から他の状態への遷移は起こらないため、一つのスイッチ指定区間に送信される音声区間は一つまでとなる。つまり、区間指定スイッチを一度押しただけで、2つ以上の発話を行ったとしても、送信されるのは最初に検出された音声区間だけとなる。ただし、発話終了後、すぐに、区間指定スイッチ4を押して、再び発話した場合に、最初の発話に対応する音声区間も二つ目の発話に対応する音声区間も送信するようにするため、状態(ii)から状態(ii)への遷移を設けておく。つまり、音声区間の信号は、入力時刻よりk時刻遅れてから送信されることを考慮して、音声区間の終端のフレームが入力されてから送信されるまでの間に区間指定スイッチ4のオンの起点が検出された場合は、状態(iii)から状態(ii)へ遷移するようにしておく。

【0084】なお、着目フレームがスイッチ指定区間として終了していないにも関わらず、再びスイッチが押されるようなことがあれば、スイッチ指定区間を延長し、再度押された区間指定スイッチ4のオフの起点からさらにLフレーム後ろまでを図9の(B)に示すように、新たなスイッチ指定区間とする。

【0085】なお、上述した実施例となる音声区間検出装置は、音声通信システムに組み込むこともできる。この場合、音声信号の送受信を行う制御部を図1の音声認識部5の代わりに設ければよい。このため、入力された音声信号の中から必要な音声区間だけを抽出し、送信することが可能となる。

【0086】

【発明の効果】本発明に係る音声区間検出装置は、音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、上記入力音声信号を一定時間分記憶する記憶手段と、上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有するので、有音声と無音声の判別による音声区間判定を常に行いながら、その判定結果と一定時間前までの入力信号を記憶しておくことで、スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出することが可能と成る。又、一つのスイッチ指定区間に対して一つの音声区間だけを検出し送信するようにしたことで、話者が発声の少し前や、少し後に関係のない言葉を発声したとしても、これを音声区間として検出することがなくなるため、例えば音声認識装置に適用した場合などに誤動作を起こしにくくなる。

17

【0087】さらに、入力信号を周波数分析して得られるパワースペクトルを用いること、及び、環境雑音の平均エネルギーに適応させてパラメータを更新することなどを行うことで、環境雑音下においても安定した音声区間検出が行えるようにすると共に、その適応化を音声区間の判定と平行して常に行っておくようにすることで、スイッチを押してから雑音環境の観測を行う必要が特になくなるため、ユーザが発話するのを待たされることもなくなる。つまり、ユーザは、単純にスイッチを押せば一回だけ発話することができるということだけを覚えておけばよく、その発話とスイッチのタイミングをそれほど意識しなくてよいものとなる。しかも、システムは、必要な音声区間だけを検出することが可能となる。

【0088】そして、この音声区間検出装置を音声認識システムや音声通信システムに組み込めば、入力信号のなかから認識処理を行うべき音声区間や送信すべき音声区間を抽出することができるようになる。

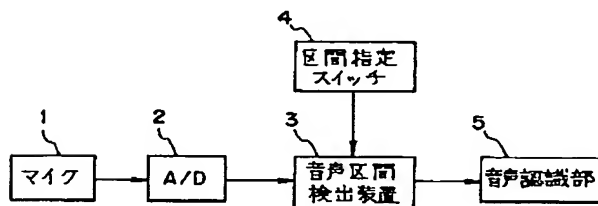
【図面の簡単な説明】

【図1】本発明の音声区間検出装置を音声認識システムに適用した例を示したブロック図である。

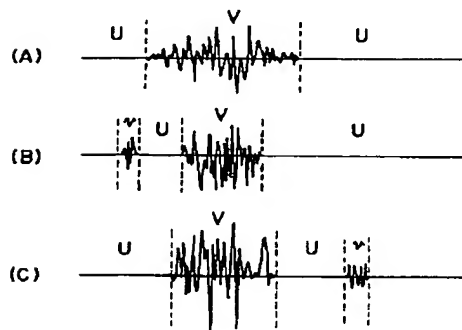
【図2】本発明の実施例の音声区間検出装置のブロック図である。

【図3】上記実施例の音声区間検出装置の動作を説明す

【図1】



【図5】



18

るためのフローチャートである。

【図4】上記実施例の音声区間検出装置の制御部の音声区間の判定について説明するための図である。

【図5】無音声部によって音声信号の一部が分離された状態を示す図である。

【図6】図5に示した分離された音声区間の判定基準を説明するための図である。

【図7】パラメータ更新における環境雑音のエネルギーと環境雑音のパラメータの関係を示した図である。

10 【図8】スイッチ信号と入力フレーム信号の時間的關係を示した図である。

【図9】スイッチ指定区間を説明するための図である。

【図10】スイッチ指定区間と音声区間の関係を示した図である。

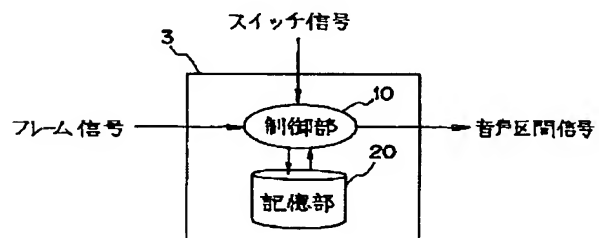
【図11】音声区間信号の送信の仕方を説明するための図である。

【図12】音声区間信号の送信の判定について示した図である。

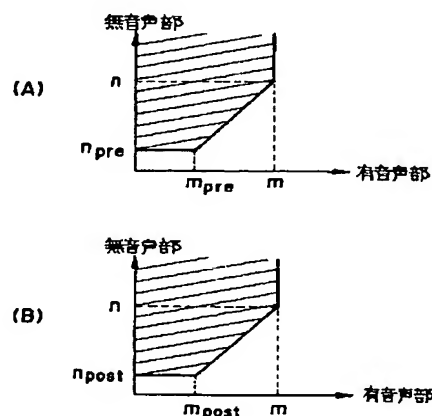
【符号の説明】

- 20 3 音声区間検出装置  
10 制御部  
20 記憶部

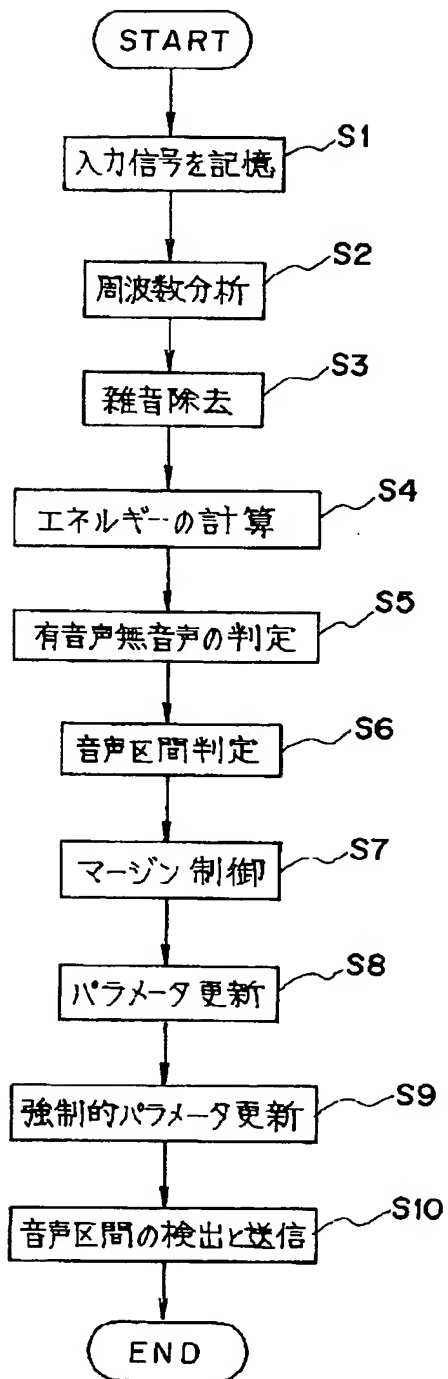
【図2】



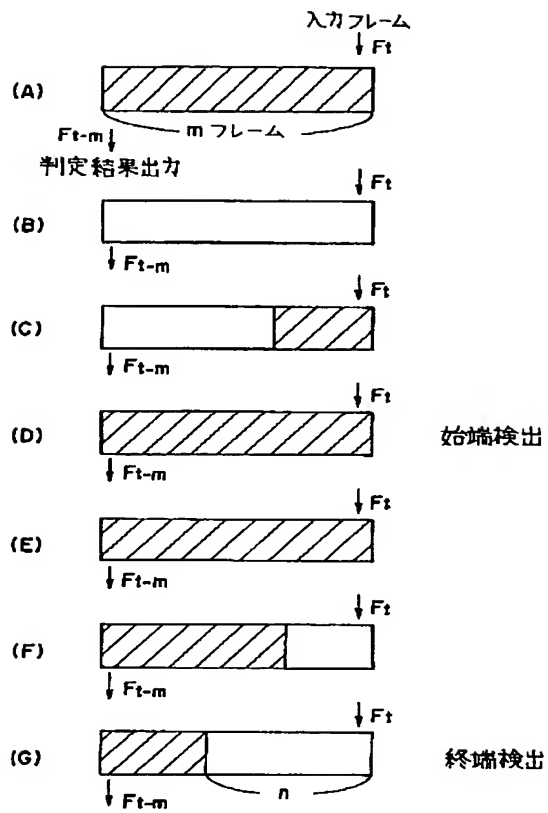
【図6】



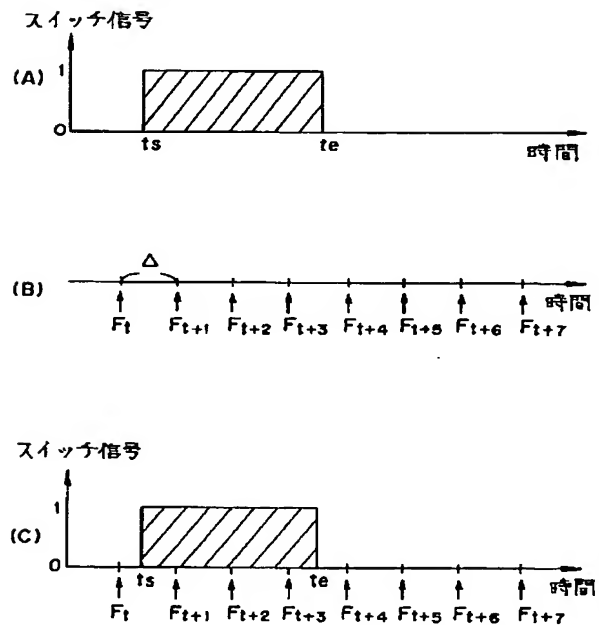
【図3】



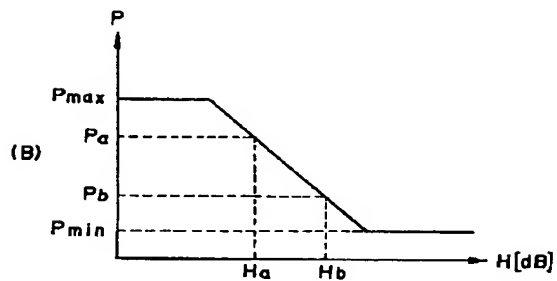
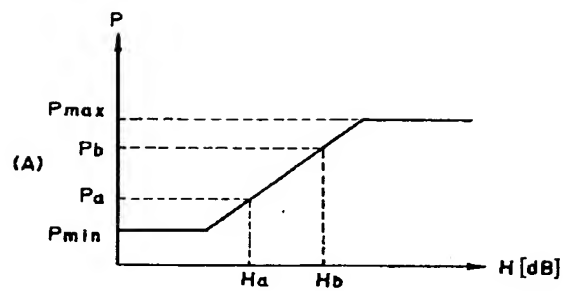
【図4】



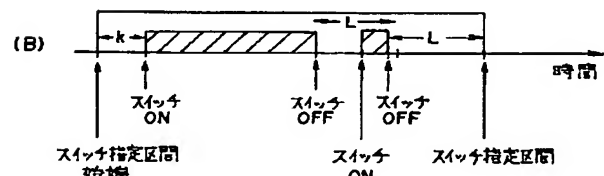
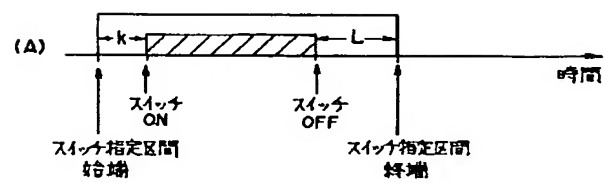
【図8】



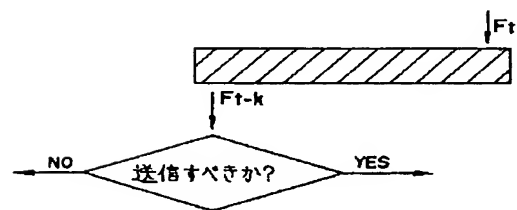
【図7】



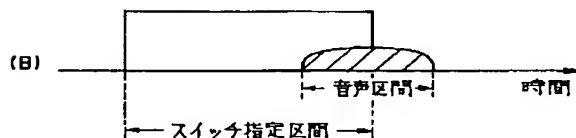
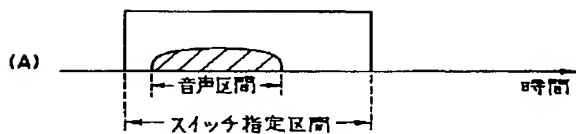
【図9】



【図11】



【図10】



【図12】

